Redundancies in Large-scale Protein Interaction Networks

Manoj Pratim Samanta and Shoudan Liang

NASA Advanced Supercomputing Division,

NASA Ames Research Center, Moffet Field, CA, USA

(Dated: June 6, 2005)

Abstract

Understanding functional associations among genes discovered in sequencing projects is a key issue in post-genomic biology [1, 2, 3, 4, 5, 6, 7]. However, reliable interpretation of the protein interaction data has been difficult [8, 9]. In this work, we show that if two proteins share significantly larger number of common interaction partners than random, they have close functional associations. Analysis of publicly available data from *Saccharomyces cerevisiae* reveals more than 2800 reliable functional associations, 29% of which involve at least one unannotated protein. By further analyzing these associations, we derive tentative functions for 81 unannotated proteins with high certainty.

A large number of genes discovered in sequencing projects remain functionally unannotated, motivating significant research in post-genomic biology. High-throughput experiments such as genome-wide monitoring of mRNA expressions as well as protein-protein interaction networks are expected to be fertile sources of information to derive their functions [1, 2, 3, 4, 5, 6, 7]. However, a high rate of false positives [8, 9] as well as the sheer volume of the data are making reliable interpretation of these experiments difficult.

In this work, we are able to overcome these difficulties by using a statistical method that forms reliable functional associations between proteins from noisy genome-wide interaction data. Our method ranks the statistical significance of forming shared partnerships for all protein pairs in the interaction network and shows that if two proteins share significantly larger number of common partners than random, they have close functional associations. In the supplement, we derive more than 2800 pairs of high quality associations for *S. cerevisiae* involving 852 proteins. The method is not overly sensitive from the false positives widely present in the two-hybrid data. Even after adding 50% randomly generated interactions to the measured dataset, we are able to recover almost all ($\sim 90\%$) of the original associations. The modular nature of the interaction network [10] is revealed by the clustering of these associations. From the derived modules, we are able to predict functions for 81 unannotated proteins with high certainty. It has been an encouraging sign that the functions of some of these proteins were recently annotated by the SGD database [11] from other sources after the completion of our work, and all but one (22 out of 23) of our predictions proved to be correct.

Our strategy of assigning statistical significance is to compare the measured protein interaction network with a random network of the same size [12, 13, 14]. The deviation of the measured network from randomness is presumed to reflect its biological significance. Non-random nature of the large-scale protein interaction network has been discussed in earlier work [12, 13, 15]. In one example, it was observed that the connectivities of the proteins in the measured interaction networks closely followed a power-law distribution instead of the exponential distribution expected from random networks [9, 12, 13, 15]. Useful biological prediction regarding the lethality of the null mutants lacking those highly connected proteins could be made from such non-random behavior [12].

We hypothesize that if two proteins have significantly larger number of common interaction partners in the measured data-set than what is expected from a random network, it would suggest close functional links between them. To validate this hypothesis, we rank all possible protein pairs

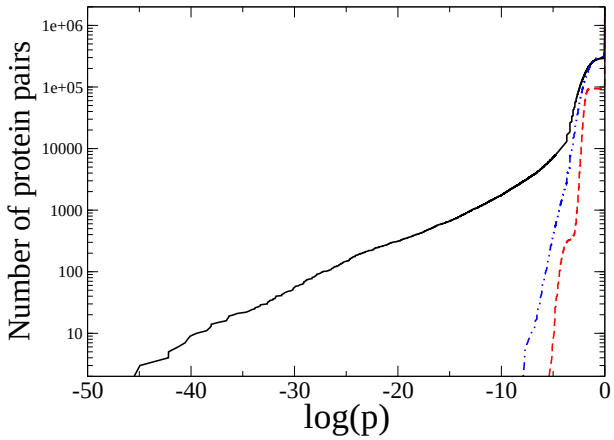


FIG. 1: Probabilities of associations for all possible protein pairs derived using our method. Solid black line: measured protein interaction network [17]; broken red line: a random network of similar size constructed by connecting randomly chosen nodes; dotted green line: a random network constructed from the measured network keeping its power-law connectivity property unchanged [13]. The probabilities of associations for the measured network are up to 40 orders of magnitude lower than the random networks.

in the order of their probabilities [16] for having the experimentally measured number of common interaction partners. If the computed probability is extremely small, it signifies that the chosen protein pair has an unusually large number of common partners. Such pairs are considered for further analysis, as we discuss in the paper.

The described method is applied on the available experimental data from budding yeast (*S. cerevisiae*) collected in the DIP database from several sources [17, 18]. In Fig. 1, we show a plot with probabilities for all protein pairs in the network sorted in increasing order. For comparison, we also show corresponding probabilities for a random network of similar size, as well as a randomized version of the measured network. The random network has the same number of nodes and edges as the measured network, but the connections are made from a uniform distribution. The randomization of the experimental network is done using a method similar to Ref. [13]. The method allows us to maintain the power-law nature of the network. As we observe from the plot, the probabilities of some of the associations in the measured network are up to 40 orders of magnitude lower than both of the randomly constructed networks. Therefore, it is safe to conclude that those associations are not artifacts due to experimental noise, but contain biologically meaningful information. It is also clear from the plot that such low probability associations did not arise from the scale-free nature of the network [12].

To understand what biological information is provided by such low-probability pairs, we inspect all pairs with probabilities below a cutoff value of 10^{-8} [19]. The detailed list is provided as a supplement [20] as well as from our website [21]. The group consists of 2833 protein pairs involving 852 proteins. A strong functional link is observed among proteins in these pairs thus validating our hypothesis. This is illustrated in Table I, where we present the ten pairs with the lowest probabilities. As we can see from the table, both proteins usually either belong to the same complex or are parts of the same functional pathway. Same trend is generally true for the larger dataset presented in the supplement. By manually inspecting the top 100 pairs, we found that in over 95% of them both proteins have similar function.

We can take advantage of the above observation to predict the functions of the unannotated proteins. About 29% of the 2833 chosen pairs contain at least one unannotated protein [22]. To assign a function to any one of them, we determine the other proteins with which it forms associations. As an example, in Table II we show that the unannotated protein $YKL059C$ shares partners with many proteins involved in transcription. Therefore, it is most likely also involved in transcription. Moreover, from the low probabilities of associations with CFT2 and CFT1, we strongly suspect that that YKL059C is involved in pre-mRNA 3' end processing. This is further confirmed by the clustering method that we present below. Our website provides an interactive tool for users to search for the close associates of any query protein and thus derive its putative function [21].

Since functionally related proteins form strong associations with each other, this can be used as the basis for an algorithm to cluster them into functional modules. We derive 202 modules [Fig. 2] from the associations and then compare the annotations of constituent proteins. 163 of the derived modules have all proteins annotated in the SGD database [11] and we find 149 of them (about 92%) to have all members of the module from the same functional complex or pathway. Therefore, if an unannotated protein belongs to the same modules with other proteins of known functions, we can predict its functions to be the same as the other ones with high confidence. By analyzing the derived modules, we predict functions for 81 unannotated proteins and present them in Table III.

We note that the chosen cut-off value (10^{-8}) is not a sharp threshold. As the number is increased, the amount of biologically meaningful information degrades gradually. In the case of the modules, their numbers and sizes increase with increasing cut-off. As an example, for the well-studied mediator complex shown in figure 2(a), as we increase the cut-off value, more proteins

Protein 1	Protein 2	Log(p)	Function
MYO3	MYO5	-47.41	Class I myosins
ROX3	SRB6	-46.12	Mediator complex
KRR1	PWP2	-45.50	snoRNA complex
ROX3	MED2	-44.94	Mediator complex
MED2	SRB6	-42.19	Mediator complex
ATP1	ATP2	-42.17	ATP complex
KAP95	SRP1	-41.25	Protein import-export
PRE1	RPN10	-40.58	Spliceosome complex
YKR081C	YNL110C	-40.33	Both unannotated
RPT1	RPN6	-40.07	Spliceosome complex

TABLE I: The ten protein pairs with the lowest probabilities [16] based on our method, along with their functions. We find both of the proteins in these pairs to belong to either the same complexes or the same functional pathways. The complete list is provided as a supplement.

known to be part of the complex come together. We find that even with cut-off as high as 2×10^{-4}, the proteins included in the mediator module are genuinely related to the complex. In our website we present an interactive program that allows users to choose different cut-off values and obtain the corresponding modules. Among the additional modules derived with higher threshold, we find two that contain mostly unannotated proteins and therefore are possibly large complexes not yet well studied by experimentalists. One of them is suspected to be involved in actin cytoskeleton organization and protein vacuolar targeting and the other one in splicing, rRNA processing and snoRNA processing. We present them in Figs. 3 and 4 expecting their identification to spur additional interest among yeast biologists.

The method presented here has several advantages. Firstly, it is not sensitive to random false positives. To illustrate, we added connections randomly increasing the average number of interactions by 50% and were still able to recover 90% of the top 2833 associations. Secondly, the method is not biased by the number of partners a protein has. As an example, JSN1, a nuclear pore protein, has the largest number of interactions in the measured dataset, but none of the 2833 associations derived by our method contains JSN1. Among the drawbacks, our method cannot extract much information about proteins with none or very few interactions in the dataset.

Associations of YKL059C	Log(p)
CFT2[T]	-32.430607
CFT1[T]	-30.151475
YSH1[T]	-28.320081
PTA1[T]	-27.843331
PAP1[T]	-27.410048
REF2[T]	-25.048611
PFS2[T]	-24.638901
YTH1[T]	-23.247919
FIP1[T]	-21.609526
HCA4[T]	-21.285573
YGR156W[U]	-17.961537
RNA14[T]	-17.732432
SWD2[U]	-14.407007
GLC7[C]	-13.284243
YOR179C[T]	-12.636400
PCF11[T]	-8.857110

TABLE II: Categories - T: transcription, U: unannotated protein, C: cellular fate/organization. Most of the associations of YKL059C are involved in transcription and therefore it is also expected to do the same. From its very low probabilities [16] of associating with CFT1 and CFT2, it is strongly suspected to be involved in pre-mRNA 3' end processing. Our website provides an interactive tool to search for the associates of any protein [21].

In conclusion, we derived functional modules and reliably predicted functions of unannotated proteins from the existence of abnormally large number of shared interaction partners in the protein-protein interaction network. We believe the real power of the method will be in studying the higher eukaryotes, where higher fraction of genes has unknown functions. Moreover, the method is applicable to other forms of networks, such as the Internet, metabolic networks, social networks and predator-prey networks.

[1] A.-C. Gavin *et al. Nature*, 415:141, 2002.

[2] Y. Ho *et al. Nature*, 415:180, 2002.

[3] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. *Proc. Natl Acad. Sci.*, 98:4569, 2001.

[4] P. Uetz *et al. Nature*, 403:623, 2000.

[5] A. H. Y. Tong *et al. Science*, 294:2364, 2001.

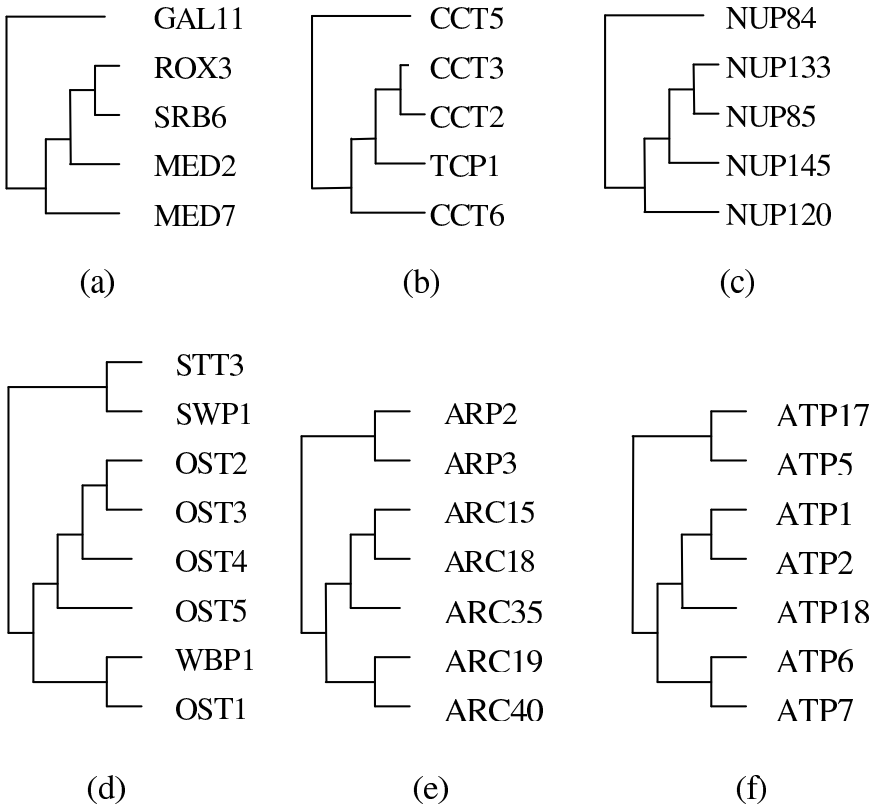


FIG. 2: Functional modules obtained by clustering the low-probability associations using an algorithm described in our paper. All proteins from each of these derived modules belong to same functional complexes. (a) Pol II transcription mediator complex, (b) chaperon ring complex, (c) nuclear pore complex, (d) oligosaccharyl transferase complex, (e) Arp2/3 complex, (f) ATP synthase complex. The complete list of modules is provided in the supplemental table.

[6] H. Zhu *et al. Science*, 293:2101, 2001.

[7] B. Grünenfelder and E. A. Winzeler. *Nature*, 3:653, 2002.

[8] P. Uetz and R. Hughes. *Curr. Opinion in Microbiol.*, 3:304, 2000.

[9] C. v. Mering *et al. Nature*, 417:399, 2002.

[10] L. H. Hartwell, J. J. Hopfield, S. Liebler, and A. W. Murray. *Nature*, 402:C47, 1999.

[11] J. M. Cherry *et al. Nucleic Acids Res.*, 26:73, 1998.

[12] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai. *Nature*, 411:41, 2001.

[13] S. Maslov and K. Sneppen. *Science*, 296:910, 2002.

[14] H. Jeong, B. Tambor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi. *Nature*, 407:651, 2000.

[15] G. D. Bader and C. W. V. Hogue. *Nature biotech.*, 20:991, 2002.

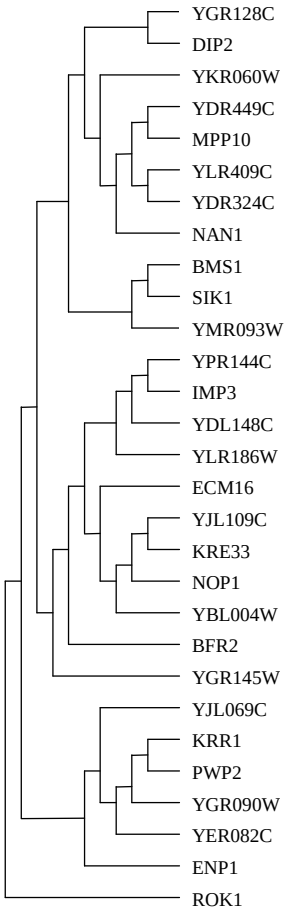


FIG. 3: A module identified by our method consisting of proteins presumably involved in assembly and maintenance of small nucleolar ribosomal complex.

[16] In a random network of N proteins, the probability that two proteins with n_1 and n_2 partners sharing m common partners is expressed as

$$P(N, n_1, n_2, m) = \frac{\begin{pmatrix} n_1 \\ m \end{pmatrix} \begin{pmatrix} N - n_1 \\ n_2 - m \end{pmatrix}}{\begin{pmatrix} N \\ n_2 \end{pmatrix}},$$

where $\begin{pmatrix} N \\ m \end{pmatrix} = \frac{N!}{m!(N-m)!}$. We start with the experimentally measured dataset of N proteins and compute the probabilities for all possible $N(N-1)/2$ pairs based on n_1, n_2 and m obtained from the experimental data. Detailed derivation of the expression is given in the supplement.

[17] C. M. Deane, L. Salwinski, Ioannis Xenarios, and D. Eisenberg. *Molecular and Cellular Proteomics*, 1.5:349, 2002.

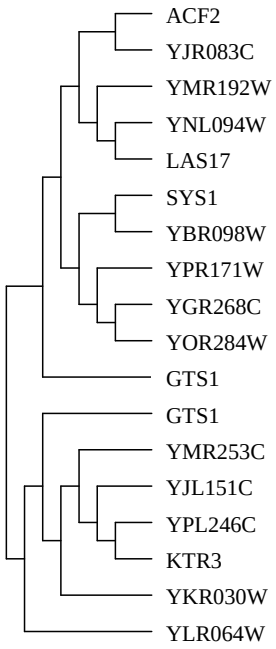


FIG. 4: A module identified by our method consisting of proteins presumably involved in actin cytoskeleton organization and protein vacuolar transport.

[18] We used 09/01/2002 update of the DIP dataset containing 14871 interactions for 4692 proteins.

[19] Since the dataset contains $N = 4692$ proteins, $1/N^2 \sim 10^{-8}$ is a reasonable cutoff. The number is validated by more rigorous comparison with random network shown in Fig. 1. However, this is not a sharp threshold as we discuss in more detail in the paper. Therefore, we present pairs up to 2×10^{-4} in the supplement.

[20] Supplemental materials are enclosed with the paper.

[21] Additional information and interactive tools are available from our website at http://www.nas.nasa.gov/bio/.

[22] Here, we use the functional classes and annotations provided in Ref. [9]. However, SGD database is constantly updating the annotations for the proteins as new information becomes available. Therefore, the actual number of unannotated proteins is lower than this source.

[23] These proteins recently got annotated in the SGD database. Except for LSG1, all other predictions proved to be correct.

[24] K. Wolfe and D. C. Shields. *Nature*, 387:708, 1997.

[25] We thank Hao Li and Ru-Fang Yeh for useful suggestions, and K. J. Harrington for critical reading of the manuscript. M. P. Samanta is supported by NASA contract DTTS59-99-D-0437/A61812D to

Protein	Predicted Function
YFR024C (YSC85), BZZ1 [23], YNL094W (APP1), YMR192W (APP2)	Actin filament organization
YGR268C (HUA1), YOR284W (HUA2), YPR171W (BSP1)	Actin patch assembly
YJR083C (ACF4)	Actin cytoskeleton organization and biogenesis
YDR036C (MRP5)	Protein biosynthesis in mitochondrial small ribosomal subunit
YKL214C (YRA2) [23]	mRNA processing/RNA metabolism
YNL207W (RIO2)	Nucleolar protein involved in 40S ribosomal biogenesis
YLR409C (UTP21), YKR060W (UTP30), YGR090W(UTP22), YER082C(UTP7) [23], YJL069C(UTP18) [23], ENP1	Associated with U3 snoRNA and 20S rRNA biosynthesis
YMR288W (HSH155) [23]	snRNA binding involved in mRNA splicing
YHR197W (IPI2), YNL182C (IPI3), YLR106C (MDN1) [23]	Ribosomal large subunit assembly and maintenance
YGR128C (UTP8) [23]	Processing of 20S pre-rRNA
YGR215W (RSM27) [23], YGL129C (RSM13) [23]	Structural constituent of ribosome
YDL213C (NOP6)	rRNA processing/transcription elongation
YNL306W (MRPS18) [23]	Mitochondrial small ribosomal subunit
YPR144C (UTP19), YDL148C (NOP14) [23], YLR186W (EMG1), YJL109C (UTP10) [23], YBL004W (UTP20)	snoRNA binding, 35S primary transcript processing
YGL099W (LSG1) [23], YDR101C (ARX1)	27S pre-rRNA ribosomal subunit
BRX1, YOR206W (NOC2), FPR1	Biogenesis and transport of ribosome
YOR145C (DIM2)	35S Primary transcript processing and rRNA modification
YEL015W (DCP3)	Deadenylation dependent decapping and mRNA catabolism
NHP10, RFX1 [23]	Modification of chromatin architecture/transcription
YDR469W (SDC1) [23]	Chromatin silencing and histone methylation
YPL070W (MUK1)	Transcription factor (or its carrier)
YLR427W (MAG2)	DNA N-glycosylase involved in DNA dealkylation
YDL076C (RXT3), YIL112W (HOS4)	Histone deacetylase complex involved in chromatin silencing
IST1	Trancription initiation factor
HCR1 [23]	Translation initiation as part of eIF3 complex
YDL074C (BRE1)	Chromosome condensation and segregation process
YGR156W (PTT1) [23], YKL059C (MPE1) [23]	mRNA cleavage and polyadenylation specificity factor
YGR089W (NNF2)	Chromosome segregation (spindle pole) and mitosis
YGL161C(YIP5) ,YGL198W (YIP4)	Vescicle mediated transport
YPL246C (QUT1), YJL151C (SNA3), YGL104C (VPS73) [23], YKR030W (MSG1)	Cell wall synthesis / protein-vacuolar targeting
YBR098W (MMS4)	Golgi to endosome transport and vescicle organization
YHR105W (YPT35)	Golgi to vacuolar transport
YBL049W (MOH1), YCL039W (MOH2)	Both same function. Possibly linked with vacuolar transport
YDL246C (SOR2)	Possibly involved in fructose and mannose metabolism
YMR322C (SNO4)	Pyridoxine metabolism
YDR430C (CYM1)	Protein involved in pyurvate metabolism
YJL199C (MBB1), YPL004C (LSP1), YGR086C (PIL1)	Metabolic protein
YLR097C (HRT3)	Nuclear ubiquitine ligase
YKR046C (PET10)	ATP/ADP exchange
YEL017W (GTT3)	Protein linked with glutathione metabolism
ITC1	Chromatin remodeling
YGR161C (RTS3)	Protein phosphatase 2A complex
EFD1	DNA replication and repair
YML117W (NAB6)	Nuclear RNA binding
YLR432W (IMD3)	RNA helicase involved in mRNA splicing
YJU2, YGR278W (CWC22), YDL209C (CWC2) [23]	Spliceosome complex involved in mRNA splicing
YGR232W (NAS6) [23], YGL004C (RPN14) [23], YLR421C (RPN13) [23]	Proteasome complex

TABLE III: Predicted functions of previously unannotated proteins.

CSC. S. Liang is supported by NASA RTOP 755-30-11.

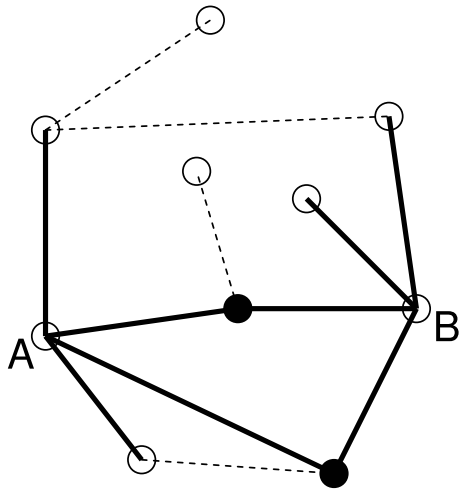


FIG. 5: In the above interaction network, both proteins A and B have 4 partners (n_1 and n_2). Two of the partners (marked by dark circles) are shared by both of them. We compute the probability for such an event to occur in a random network. If the computed probability is low, perhaps two proteins are redundant in their functions.

APPENDIX A: SUPPLEMENTAL INFORMATIONS (THEORY)

1. Methods

a. Mathematical Expression for Probability

In a network of N proteins, the probability that two proteins with n_1 and n_2 partners [Fig. 5] share m common partners is given by

$$P(N, n1, n2, m) = \frac{\begin{pmatrix} n_1 \\ m \end{pmatrix} \begin{pmatrix} N - n_1 \\ n_2 - m \end{pmatrix}}{\begin{pmatrix} N \\ n_2 \end{pmatrix}} = \frac{(N-n_1)!(N-n_2)!n_1!n_2!}{N!m!(n_1-m)!(n_2-m)!(N-n_1-n_2+m)!} \tag{A1}$$

The above expression is symmetric with respect to interchange of n_1 and n_2. Eq. A1 is derived in the following manner. It is a ratio where the denominator is the total number of ways two proteins can have n_1 and n_2 partners given by

$$\begin{pmatrix} N \\ n_1 \end{pmatrix} \begin{pmatrix} N \\ n_2 \end{pmatrix}, \tag{A2}$$

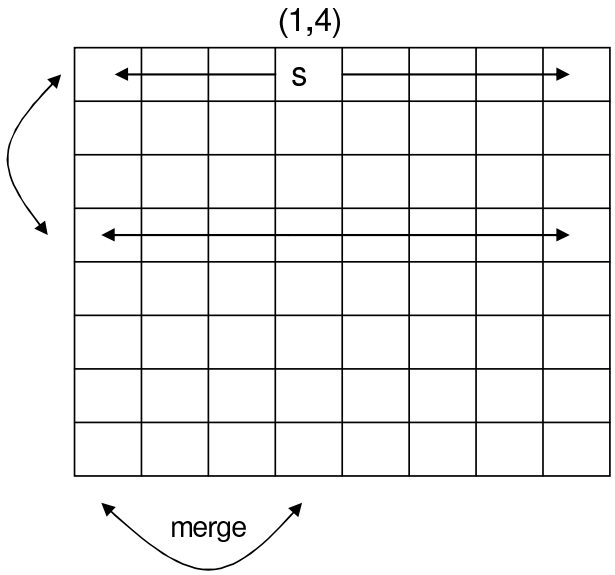


FIG. 6: In our clustering algorithm, we start with a matrix with p-values for all pairs. If the element (m, n) has the lowest p-value, a cluster is formed with proteins m and n. Therefore, rows/columns m and n are merged with new p value of the merged row/column as geometric mean of the separate p values of the corresponding elements.

whereas the numerator is the number of cases among them where m of those partners are common to both of them. It is expressed as

$$\binom{N}{m}\binom{N-m}{n_1-m}\binom{N-n_1}{n_2-m} = \binom{N}{n_1}\binom{n_1}{m}\binom{N-n_1}{n_2-m}. \tag{A3}$$

The numerator can be derived using the following argument. In the combinatorial product on the left hand side, the first term represents the number of ways m common partners can be chosen from all N proteins. For the first protein, we choose $n_1 - m$ remaining partners out of remaining $N - m$ proteins. This is the second term in the product. Subsequently for the second protein, we choose $n_2 - m$ remaining partners none of which match any of n_1 partners of the first protein. This contributes to the third term.

For the calculations in our paper, the results are approximately the same, whether we compute the probabilities for pairs with exactly m common partners or we compute for m or more partners. It can be checked from the expression of probability in Eq. A1, that probability terms for increasing m fall inversely with N. Since N for our case is about 5000, the additional terms in the probability expression are negligible.

Protein 1	Protein 2	Index
MYO3	MYO5	1
GIC1	GIC2	72
TIF4632	TIF4631	145
NUP100	NUP116	476
HSC82	HSP82	485
ZDS1	ZDS2	564
PPH21	PPH22	579
KCC4	GIN4	606
RFC3	RFC4	634
CLN1	CLN2	918
GSP2	GSP1	1288
YPT32	YPT31	1550
BOI1	BOI2	1640
SEC4	YPT7	1785
YPT53	VPS21	1888
BMH1	BMH2	1920
PCL7	PCL6	1926
YGR010W	YLR328W	2162
MYO4	MYO2	2474
SAP190	SAP185	2721
MKK1	MKK2	2725
IMD4	YLR432W	2746

TABLE IV: Associations derived by us which were also ancient paralogs according to Ref. [24]. Third column in the table represents the indices for the pairs in the list of associations sorted according to increasing probabilities. The list is also available as a supplementary material.

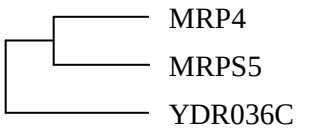
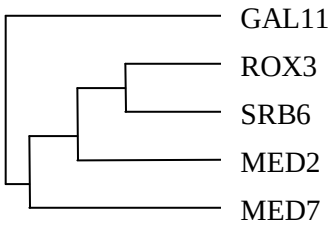
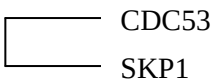
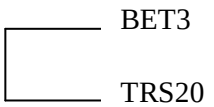
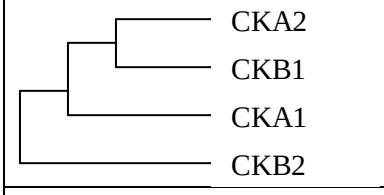
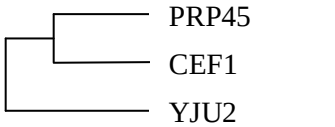
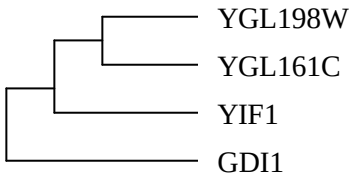
b. Clustering Technique

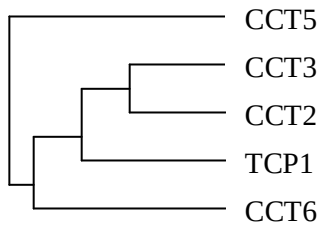
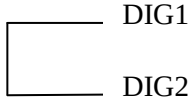
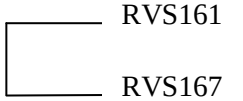
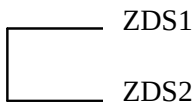
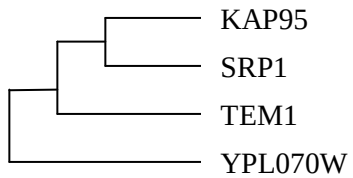
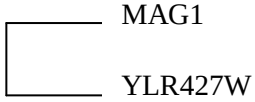
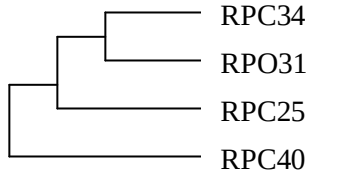
Our clustering method is as follows. We compute p values for all possible protein pairs and store them in a matrix. Then we pick the protein pair with lowest p value and choose it as the first group in the cluster. The rows and columns for these two proteins are merged into one row and one column [Fig. 6]. Probability numbers for this new group are geometric means of the two probabilities [or arithmetic means of the $log(p)$ values]. The process is continued repeatedly, thus adding more and more clusters as well as making the existing ones bigger, until a threshold is reached.

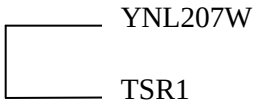
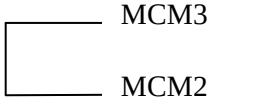
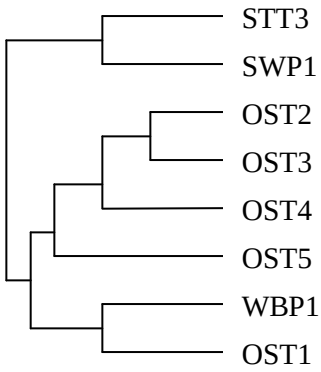
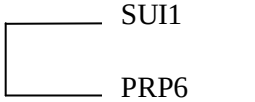
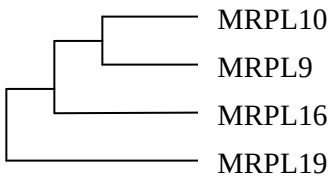
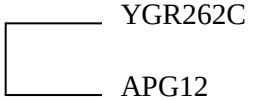
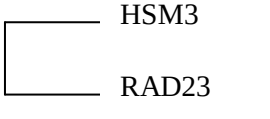
2. Ancient Paralogs

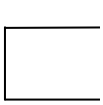
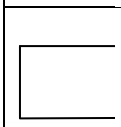
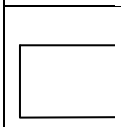
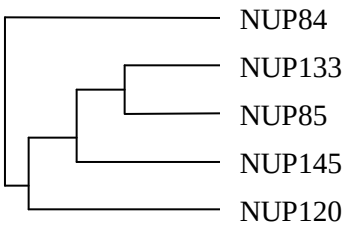
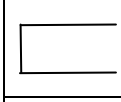
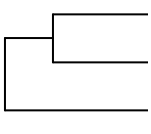
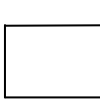
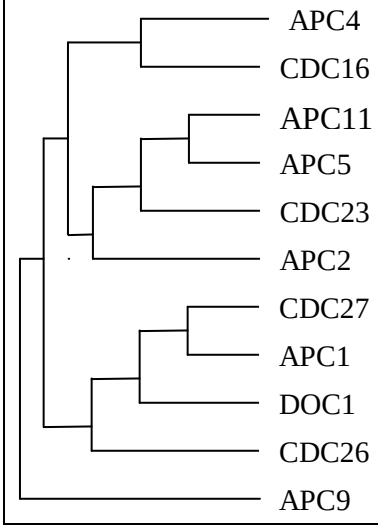
Ref. [24] proposed possibility of duplication of the entire yeast genome in some distant past and presented a list of genes that were identical or matched closely due to this event. We check how many of the associations derived by us were also such ancient paralogs and present them in Table. IV. We find 22 such ancient paralogs among the list of top 2833 pairs (.7%). Therefore, these are the ancient paralogs that maintained their functions over time.

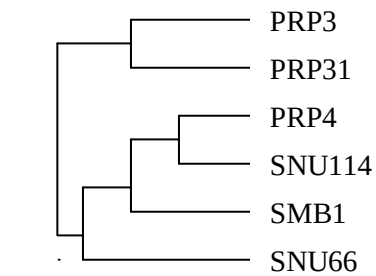
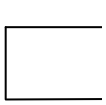
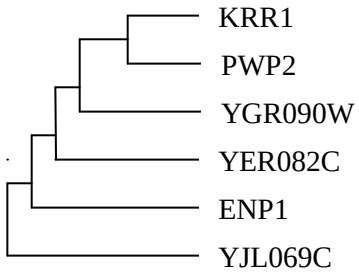
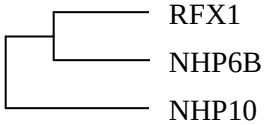
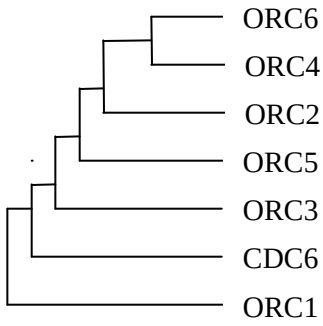
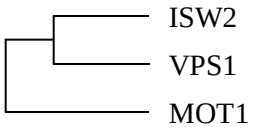
APPENDIX B: SUPPLEMENTAL INFORMATIONS (FUNCTIONAL MODULES DERIVED USING OUR TECHNIQUE)

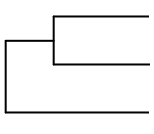
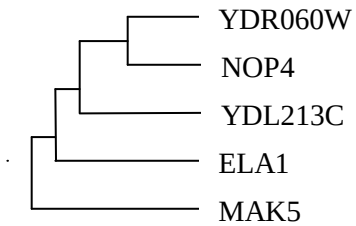
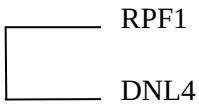
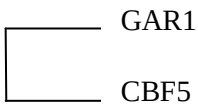
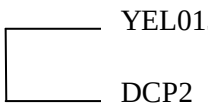
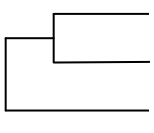
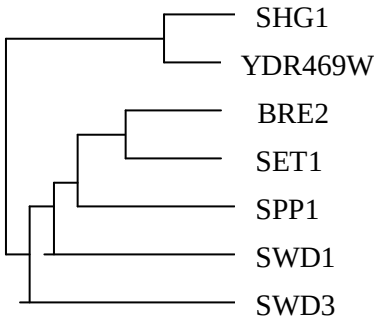
	MRP4 and MRPS5 are structural constituents of mitochondrial small ribosomal subunit involved in protein biosynthesis. Therefore, yet un-annotated YDR036C (MRP5) is strongly suspected to have similar function. Their functional link can be checked by double or triple deletion experiments.
	Parts of RNA polymerase II transcription mediator complex involved in transcription from Pol II promoter.
	Parts of nuclear ubiquitin ligase complex involved in ubiquitin-protein ligase during G1/S and G2/M transitions of mitotic cell cycle.
	Parts of TRAPP complex involved in targeting and fusion of ER to Golgi vesicles.
	Alpha and beta subunits of casein kinase II complex involved in regulation of several cellular processes.
	PRP45 and CEF1 are involved in pre-mRNA splicing with the spliceosome. Therefore, yet un-annotated YJU2 is strongly suspected to have similar function.
	GDI1 is a RAB GDP-dissociation inhibitor involved in vesicle-mediate transport, whereas YIF1 is part of COPII-coated vesicle involved in ER to Golgi transport. Therefore, yet unannotated YGL161C (YIP5) and YGL198W (YIP4) are strongly suspected to be involved in similar cellular functions. YGL161C and YGL198W are not homologous to remaining two proteins.

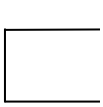
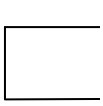
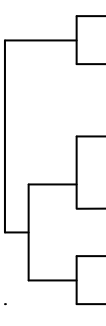
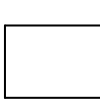
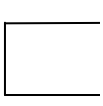
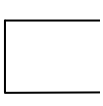
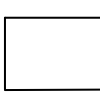
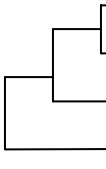
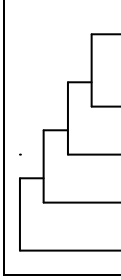
	Parts of chaperone ring complex located in the cytoplasm and assisting in protein folding.
	Both MAP kinase-associated proteins involved in down-regulation of invasive growth. It is known that each of them is viable, but double deletion results in loss of function.
	Parts of actin cortical patch linked with endocytosis. Close functional link is well known.
	Both control cell-polarity, although ZDS1 is located in bud-tip and ZDS2 at the nucleus. It is known that each gene is viable but double deletion affects cell-cycle progression. They are also ancient paralogs.
	KAP95 is part of nuclear pore complex involved in protein nuclear import. SRP1 is part of nuclear pore complex involved in nucleo-cytoplasmic transport. TEM1 is a GTP-binding protein involved in termination of M-phase of cell cycle. Link is not clear to us. These four protein also have strong association with GCD7, a translation initiation factor. We suspect YPL070W (MUK1) is either a transcription factor or involved in transporting of transcription factor.
	MAG1 is an alkyl-base DNA N-glycosylase involved in DNA dealkylation. Therefore, YLR427W (MAG2) is suspected to have similar function.
	Parts of Pol III complex involved in transcription from Pol III promoter. In addition, RPC40 is also part of Pol I complex.

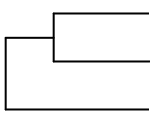
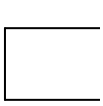
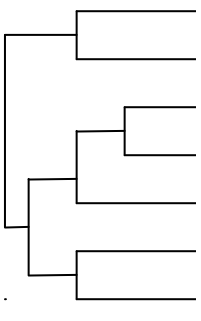
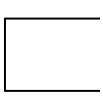
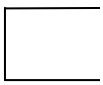
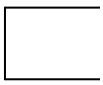
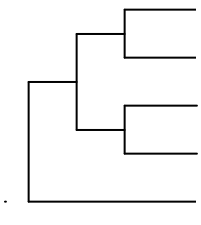
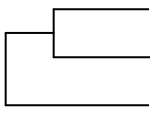
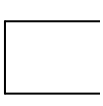
	TSR1 is a nucleolar protein involved in 40S ribosomal biogenesis. Therefore, yet unannotated YNL207W (RIO2) possibly have similar function. Moreover, both of them share common partners with ENP1, a nuclear protein linked with cell-growth and maintenance.
	Both ATP-dependent helicases involved in DNA-replication initiation, DNA unwinding and pre-replicative complex formation.
	Parts of oligosaccharyl transferase complex involved in N-linked glycosylation.
	SUI1 is part of translation initiation factor whereas PRP6 is a pre-mRNA splicing factor.
	Mitochondrial ribosomal protein of the large subunit.
	APG12 is a membrane-located protein involved in autophagy and protein-vacuolar targeting, whereas YGR262C (BUD32) is a protein serine/threonine kinase linked with bud site selection. We suspect they have closer functional link. Both of them share many partners with the ubiquitin complex.
	HSM3 is involved in DNA mismatch repair pathway, whereas RAD23 is involved in nucleotide-excision repair and DNA damage recognition.

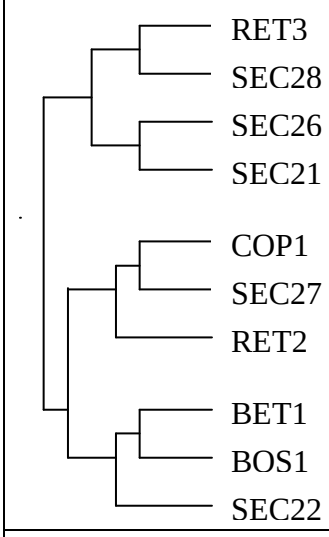
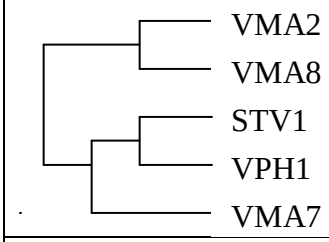
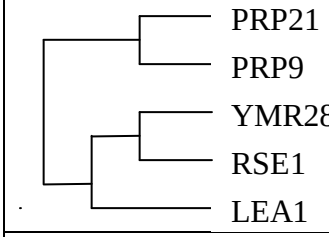
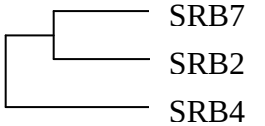
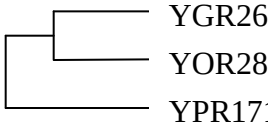
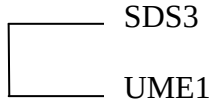
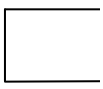
	Regulation of cell-cycle controlling the START point. They are also ancient paralogs.
	SOR1 is involved in fructose and mannose metabolism. Therefore, yet unannotated YDL246C (SOR2) also possibly a metabolic protein.
	Parts of cohesion complex involved in mitotic sister chromatid cohesion.
	Forms a nuclear pore complex importing-exporting materials between nucleus and cytoplasm.
	Proteins involved in organizing actin filaments for cell polarization and endocytosis.
	Translation initiation factor. TIF4632 and TIF4631 are ancient paralogs.
	Regulators of silencing at HML, HMR, and telomeres.
	Anaphase-promoting complex involved in mitotic metaphase anaphase transition.

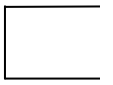
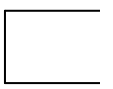
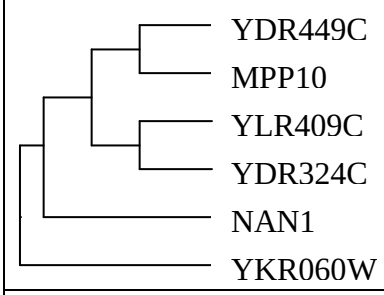
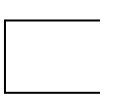
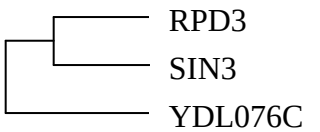
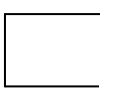
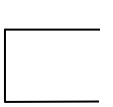
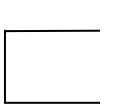
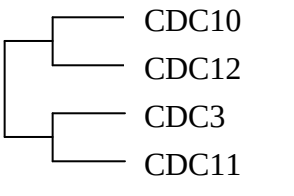
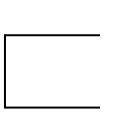
	Involved in pre-mRNA splicing.
	Involved in cytoskeleton organization and biogenesis.
	Linked with snoRNA complex involved in processing of 20S pre-rRNA.
	SEH1 is a nuclear-pore protein involved in import-export between nucleus and cytoplasm. SEC13 is a protein involved in release of transport vesicles from the ER nuclear pore complex subunit.
	NHP6B is a chromatin binding protein involved in establishment and maintenance of chromatin architecture as well as regulation of transcription from Pol II and Pol III promoters. Therefore, other two are strongly suspected to have similar function. We note that RFX1 has DNA binding domains.
	Origin recognition complex involved in DNA replication.
	ISW2 is an ATPase linked with chromatin modeling. MOT1 is an ATPase regulating transcription from Pol II promoter. VPS1 is a GTPase involved in protein-vacuolar targeting and vacuolar transport. It is not clear to us why VPS1 forms strong associations with the other two.

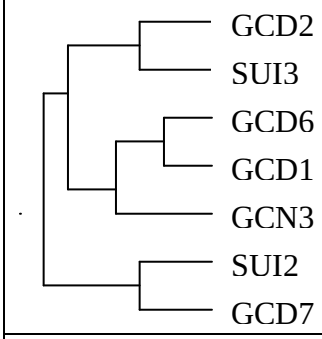
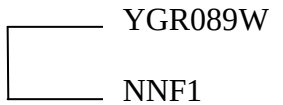
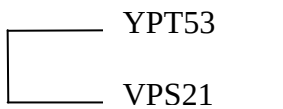
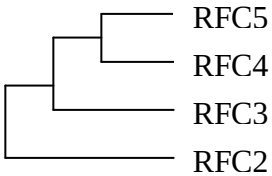
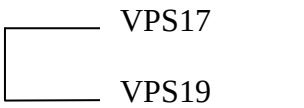
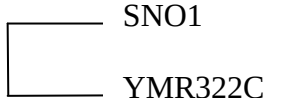
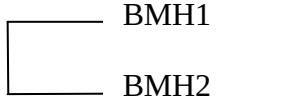
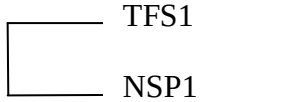
	Pol II transcription factors.
	MAK5, NOP4 and YDR060W are involved in rRNA processing and ribosomal large subunit assembly and maintenance. ELA1 is a transcription elongation factor involved in RNA elongation from Pol II promoter. The link of it with the other three is not clear to us. Unannotated YDL213C (NOP6) is strongly suspected to have similar function.
	Both RPF1 and DNL4 form strong associations with other proteins involved in ribosomal large subunit assembly and maintenance. DNL4 is currently annotated as a DNA ligase that is active in double strand break repair via non-homogeneous end joining.
	35S primary transcript processing in small nuclear ribo-nucleoprotein complex.
	DCP2 is linked with deadenylation-dependent decapping and mRNA catabolism. Therefore, YEL015W (DPC3) is suspected to be involved in similar functions. They both share many common partners with LSM proteins of small nuclear ribo-nucleoprotein complex.
	They all belong to stationary phase-induced gene family involved in pyridoxine metabolism.
	Complex involved in chromatin silencing at telomere and histone methylation.

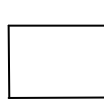
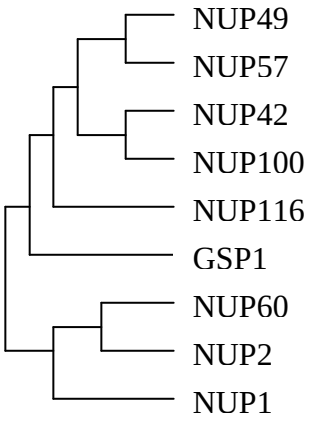
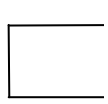
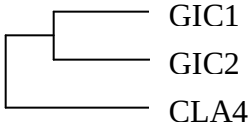
	TFIID complex active with transcription initiation from Pol II promoter.
	Involved in DNA-replication initiation.
	Arp2/3 complex involved in cell growth and maintenance.
	CDC28-complex involved in the cell-cycle.
	They are both yeast RNA Annealing proteins. YRA1 is linked with mRNA processing. Therefore, we suspect YKL214C (YRA2) has similar function.
	Both involved in homotypic vacuole fusion (non-autophagic) vacuole organization and biogenesis. Since they are both viable, double deletion experiment can be tried.
	They are both heat-shock proteins involved as chaperone helping protein folding. SSA1 is also involved in protein transport between nucleus and cytoplasm. It is grouped with SRM1 complex in cluster-1069.
	YSC84 and SLA1 are involved in actin filament organization. Therefore, unannotated YFR024C and BZZ1 are strongly suspected to be involved in similar functions.
	Parts of histone deacetylase complex.

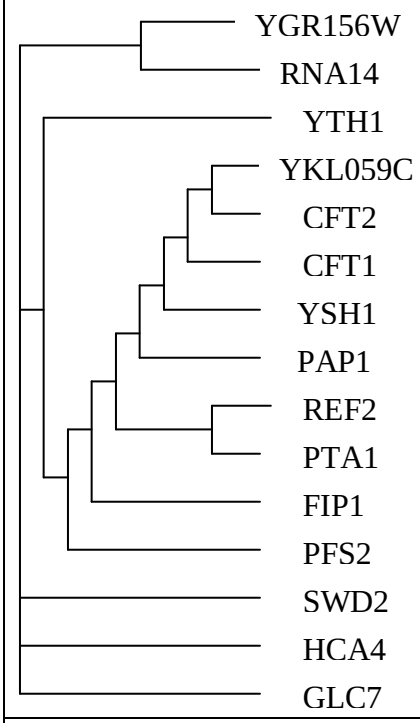
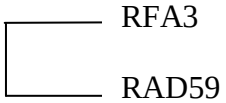
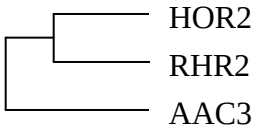
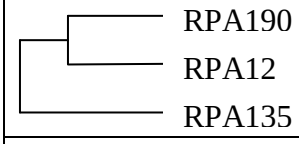
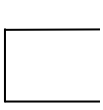
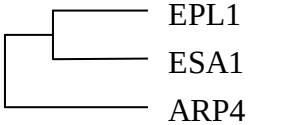
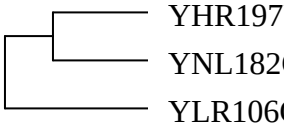
	Parts of COPII-coated vesicle complex linked with ER to Golgi transport.
	Involved in rho-protein signal transduction and establishment of cell polarity.
	Proton-transporting ATP synthase complex.
	Transcription factor TFIIF large subunit.
	SEC7 is a ARF guanyl-nucleotide exchange factor involved in protein transport, whereas LCB2 is a serine palmitoyltransferase involved in sphingolipid biosyntheis. Link is not clear to us.
	TRAPP involved in ER to Golgi transport.
	DNA-directed RNA polymerase II, core.
	Protein phosphatase type 2A complex. PPH21, 22 are ancient paralogs.
	Two unannotated proteins that form associations with each other. From their other associations, we suspect that they may be involved in vacuolar transport. Since they are both viable, double deletion experiment will possibly link them and identify their function.

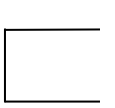
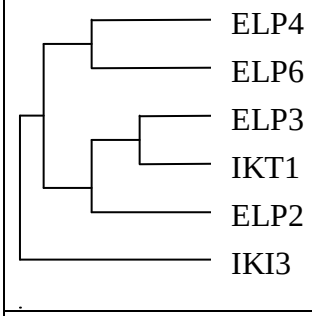
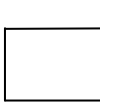
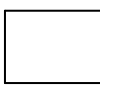
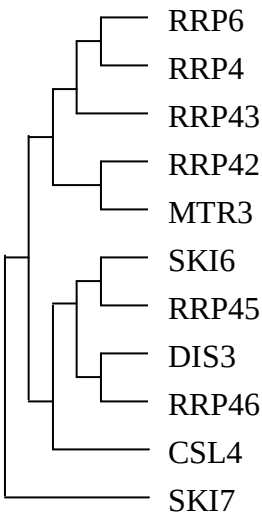
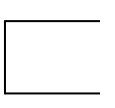
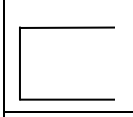
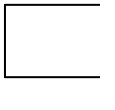
	COPI or COPII vesicle coats involved in ER to Golgi transport or retrograde transport.
	Hydrogen transporting ATPase.
	U2 snRNA binding involved in mRNA splicing. Unannotated YMR288W (HSH155) is strongly suspected to be involved in the same process.
	Suppressor of RNA polymerase II, possible component of the holoenzyme. They share many partners with members of the mediator complex.
	They are unclassified proteins strongly suspected to be involved with actin patch/filament assembly aiding in cytokinesis. They also share many partners with members of the ACF2-complex discussed below.
	UME1 is a transcription factor. SDS3 is part of histone deacetylase complex involved in transcriptional gene silencing.
	SMC4 is involved in mitotic chromosome condensation. NUM1 is involved in polymerization and stabilization of microtubules.

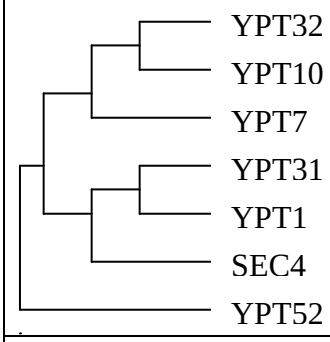
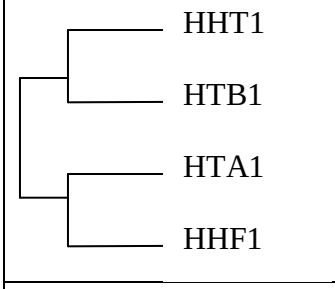
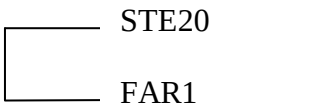
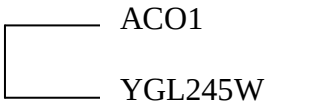
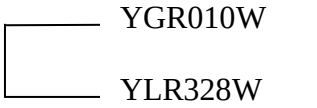
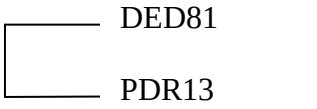
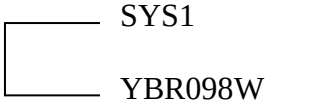
	Peroxisome organization and biogenesis.
	ARG80 is a transcription factor regulating genes ARG81 and ARG82. MCM1 is also a transcription factor.
	NAN1, YDR324C, YDR449C and MPP10 are linked with snoRNA binding and processing of 20S pre-rRNA. Therefore, the other un-annotated are linked with the same function.
	SEN15 is a tRNA-intron endonuclease complex involved in tRNA splicing. HRR25 is a casein kinase linked with several biological processes. The link between two is not clear. HRR25 also shares many partners with TEM1.
	RPD3 and SIN3 are part of histone deacytalase complex involved in chromatin silencing. Therefore, we strongly suspect YDL076C (RXT3) to be involved in the same function.
	Both proteins are involved in protein folding.
	Both proteins regulate actin cytoskeleton. Double deletions are temperature sensitive and show deficiency in bud emergence.
	GCD11 is a translation initiation factor. Therefore, we strongly suspect unannotated IST1 to be involved in similar process. It forms associations with many other proteins acting as translation initiation factors.
	They all are involved in proper bud growth. Only CDC10 is viable. They share partners with proteins of KCC4 group involved in bud ring formation.
	KAP120 is structural constituent of nuclear pore, whereas SSA2 is involved in chaperoning as well as SRP-dependent, membrane targeting, translocation. Double deletion should be studied. They also share partners with other nuclear pore complex containing NUP133 as well as transport proteins involving MTR10.

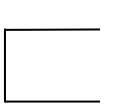
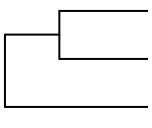
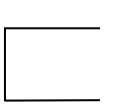
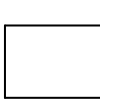
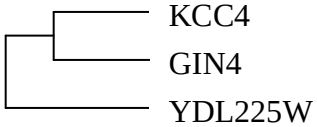
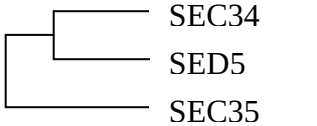
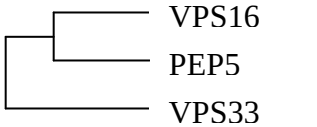
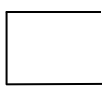
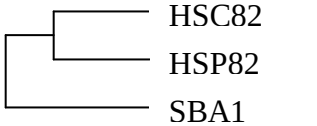
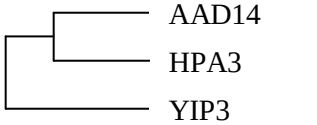
	Eukaryotic translation initiation factor 2 (eIF2) complex.
	NNF1 is involved in chrosome segregration (spindle pole) and mitosis. Therefore, we suspect un-annotated YGR089W (NNF2) is involved in similar function.
	Both are GTPase proteins linked with endocytosis in late endosome. Double deletion experiment is recommended.
	DNA replication factor C complex. RFC3 and RFC4 are ancient paralogs.
	Both proteins are involved in vacuole to golgi or endosome to golgi transport. Double deletion recommended.
	SNO1 is involved in pyridoxine metabolism. Therefore it is likely YMR322C (SNO4) is involved in same function. Double deletion is recommended.
	Both are RAS signaling proteins that activate MAPK. It is known that null mutants of individual genes are viable but double deletion is inviable.
	NSP1 is a nuclear pore protein involved in transport. Unannotated TFS1 is a lipid-binding protein. Therefore, it is possibly involved in transport between nucleus and cytoplasm using the NSP1 pore.

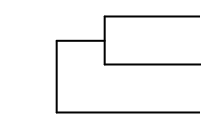
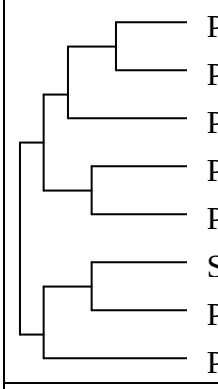
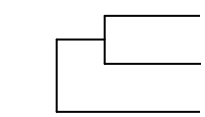
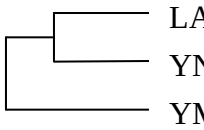
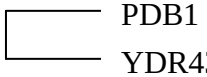
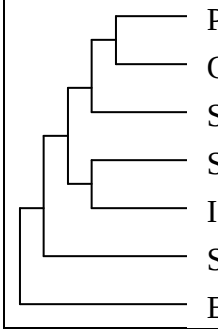
	They are both cyclin-dependent protein kinases acting as regulator proteins. Double deletion is recommended.
	Nuclear pore proteins involved in NLS-bearing substrate-nucleus import, mRNA-binding (hnRNP) protein-nucleus import, mRNA-nucleus export, nuclear pore organization and biogenesis, protein-nucleus export, rRNA-nucleus export, ribosomal protein-nucleus import, snRNA-nucleus export, snRNP protein-nucleus import, tRNA-nucleus export. NUP100 and NUP116 are ancient paralogs.
	HRT1 belongs to nuclear ubiquitin ligase complex involved in cell-cycle. Therefore, YLR097C (HRT3) is possibly linked with similar function.
	Active in rho-protein signal transduction pathway. Also forms associations with ZDS1, ZDS2 controlling cell-polarity. GIC1 and GIC2 are ancient paralogs.

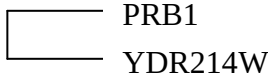
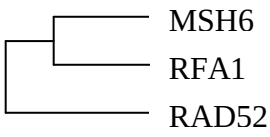
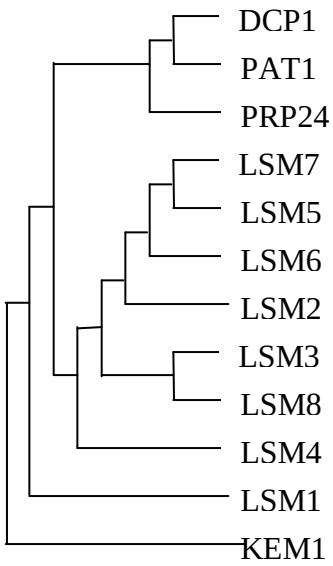
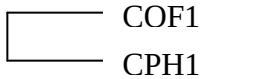
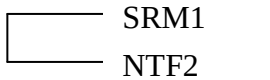
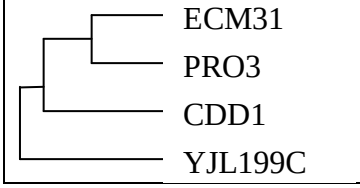
	They are parts of mRNA cleavage and poly-adenylation specificity factor complex. HCA4 is linked with 35S primary transcript processing.
	Involved in DNA repair.
	All metabolic proteins. HOR2 and RHR2 are involved in glycerol metabolism and response to osmotic stress. AAC3 is involved in ATP/ADP exchange.
	DNA-directed RNA polymerase I complex.
	Both involved in metabolic process.
	Regulation of transcription from Pol II promoter as histone acetyltransferase complex.
	They are linked with a large group of unannotated proteins linked with ribosomal large subunit assembly and maintenance.

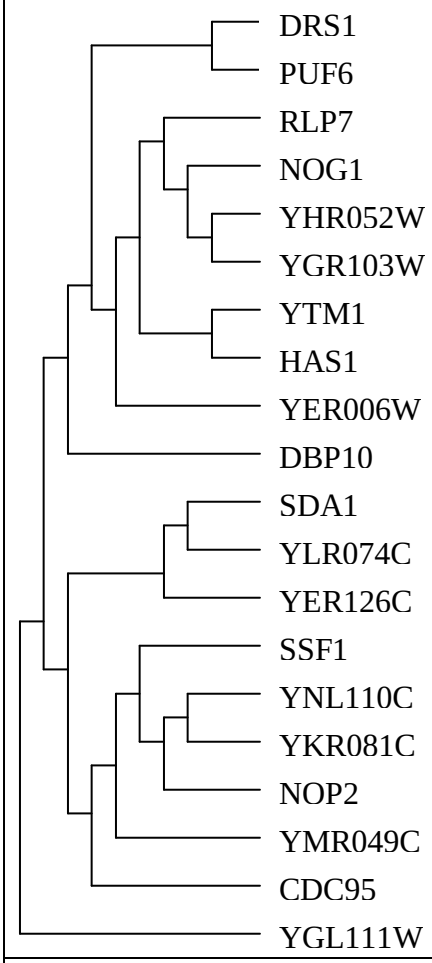
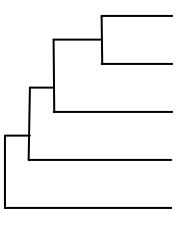
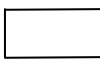
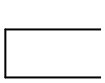
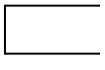
	Both involved in snoRNA binding and 35S primary transcript processing.
	Transcription elongation factor complex regulating of transcription from Pol II promoter.
	GCN2 is a protein kinase linked with protein amino acid phosphorylation, whereas NHP2 is associated with 35S primary transcript processing.
	TFIID complex, involved in general RNA polymerase II transcription factor.
	They all form nuclear exosome (RNase complex) and are involved in 35S primary transcript processing and mRNA catabolism.
	Processing of 20S pre-rRNA.
	They are both structural constituents of ribosome involved in protein biosynthesis. Double deletion strongly recommended.
	YMR145C is a NADH dehydrogenase involved in ethanol fermentation, whereas BGL2 is a glucan 1,3 beta-glucosidase involved in cell wall organization and biogenesis. Their link is not clear to us. Double deletion recommended.

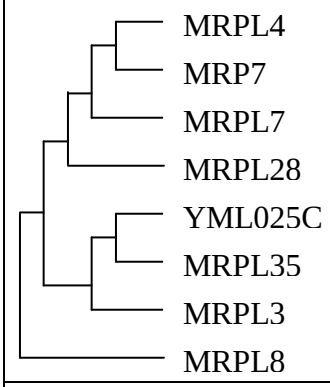
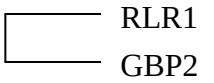
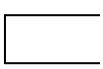
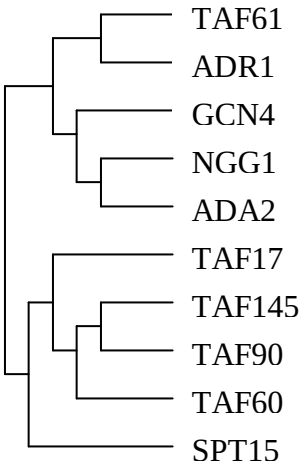
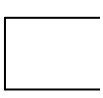
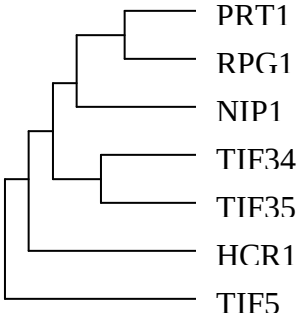
	They are involved in Golgi to vacuole transport or vesicle-mediated transport.
	They are parts of nucleosome complex involved in chromatin assembly and disassembly.
	STE20 is a protein kinase involved in cell cycle progression, whereas FAR1 is a protein kinase inhibitor involved in cell-cycle arrest.
	ACO1 is a aconitate hydratase involved in *glutamate* biosynthesis, whereas YGL245W is a *glutamate*-tRNA ligase.
	Both involved in nicotinamide adenine dinucleotide metabolism. Double deletion recommended.
	SNF1 is a protein kinase linked with glucose metabolism, whereas SNF4 is a protein kinase activator involved in regulation of transcription from Pol II promoter. Possibly they are in the same pathway.
	DED81 is a asparagine-tRNA ligase, and PDR13 is involved in protein biosynthesis and chaperone. They are also linked with other proteins involved in glycine-tRNA aminoacylation (ERG10 and GRS1).
	FOL2 is a GTP cyclohydrolase involved in folic acid and derivative biosynthesis, whereas POR1 is involved in ion transport and aerobic respiration.
	SYS1 is involved in golgi to endosome transport and vescicle organization and biogenesis. It is not clear why YBR098W which also shares many partners with other proteins involved in cellular organization is classified as DNA repair protein in SGD.

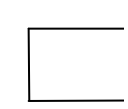
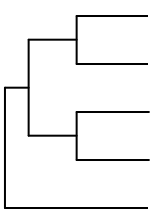
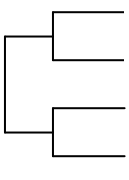
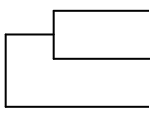
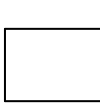
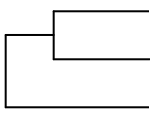
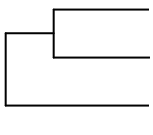
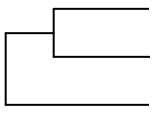
	They both belong to respiratory chain complex III and involved in aerobic respiration.
	Ribosomal biogenesis and pre-rRNA processing.
	They are both MAP kinases involved in signal transduction of mating signals. Double deletion should be interesting.
	Involved in Rho protein signal transduction, actin filament organization, establishment of cell polarity, polar budding and pseudohyphal growth.
	YDL225W is involved in cytokinesis and formation of bud-ring. The remaining two are protein kinases active in axial budding, bud growth, protein amino acid phosphorylation, septin assembly, septum formation, septin checkpoint at the bud neck. KCC4 and GIN4 are ancient paralogs.
	Involved in ER to Golgi transport and intra-Golgi transport.
	Involved in Golgi to endosome transport, homotypic vacuole fusion (non-autophagic), late endosome to vacuole transport, protein-vacuolar targeting and vacuole organization and biogenesis.
	Cyclin-dependent protein kinase involved in mitotic induction.
	HSC82 and HSP82 are heat shock proteins, ancient paralogs, whereas SBA1, a protein linked with chaperoning, is known to bind with HSP90 heat shock complex.
	HPA3 - Histone acetylation, AAD14 - aldehyde metabolism and YIP3-COPII-coated vesicle. Link unclear. Synthetic mutation recommended. Most of their other partners are involved in different metabolic functions.

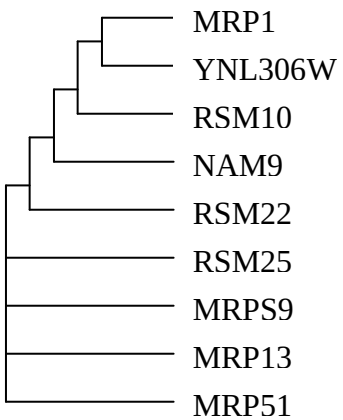
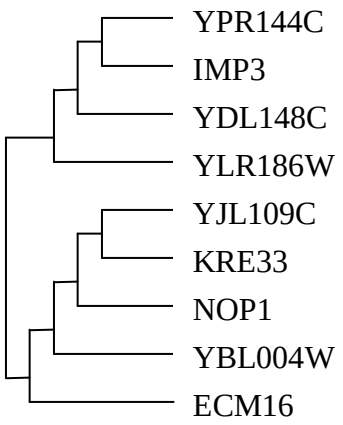
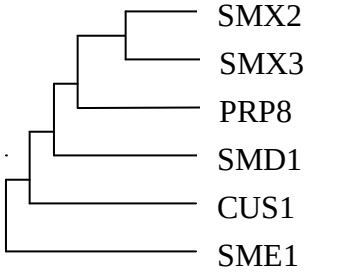
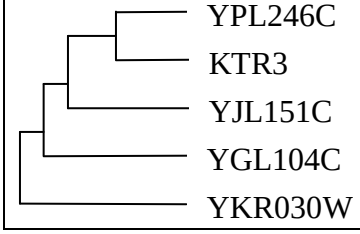
	Both MYO3 and MYO5 are class I myosins involved in cytokinesis through transport of membrane bound components. Ancient paralogs originating from ancient gene duplication. Deletion of either of them has little effect on cell growth, but double deletion causes severe defects in growth and actin cytoskeleton organization. Link with UBP7 is not clear.
	They are part of 20S core proteasome involved in ubiquitin dependent protein catabolism.
	They are involved in vacuolar acidification. Double or triple mutation should be tried.
	THI4 is involved in thiamin biosynthesis and DNA repair. Therefore, YNK1, a nucleoside diphosphase kinase is suspected to be linked in the same pathway. Double deletion study will clarify their link.
	Actin filament assembly. YMR192W (APP2) is unannotated. YNL094W (APP1) is partly annotated linked with actin filament assembly.
	PDB1 is a pyurvate dehydrogenase involved in pyurvate metabolism. Therefore, yet unannotated YDR430C (CYM1), a cystolic metalloprotease is suspected to have similar function. Double deletion experiment is recommended.
	Splicosome complex involved in mRNA splicing.

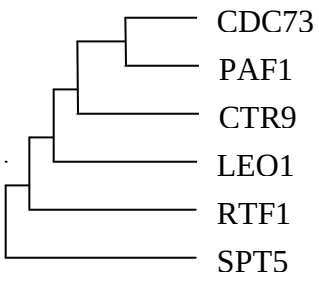
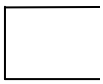
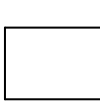
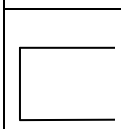
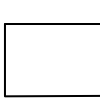
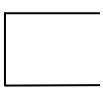
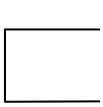
	PRB1 responds to starvation and needed for full protein degradation during sporulation. YDR214W (AHA1), a heat shock protein is suspected to have similar function in shock response. Double deletion study is recommended. (low confidence)
	They are all involved in DNA strand annealing and repair.
	Parts of snRNA complex involved in mRNA splicing.
	COF1 is linked with actin filament depolymerization whereas CPH1 is associated with histone deacetylase complex. Their other links (OYE2, CYR1, MYO4 etc.) make us suspect that these two processes are linked with each other.
	Signal-transducer and nucleus-cytoplasm transport.
	CDD1-cytidine deaminase, ECM31-pantothenate biosynthesis, PRO3-proline biosynthesis, therefore, YJL199C (MBB1) is likely to be involved in metabolic process.

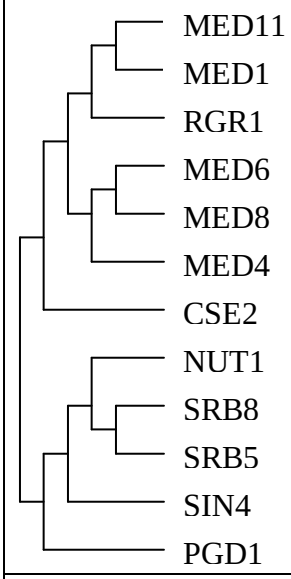
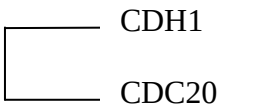
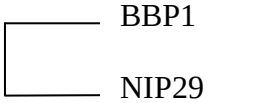
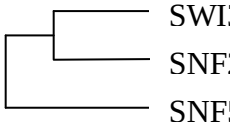
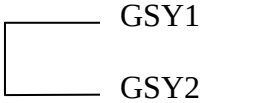
	Ribosomal large subunit assembly and maintenance.
	25kDa RNA-polymerase subunit common to all Pol I, II and II.
	VAM7 is v-SNARE protein linked with Golgi to vacuole transport. Therefore, we strongly suspect that YHR105W (YPT35) is active in similar function. It also shares partners with YPT1 and YPT32 involved in similar functions. Double deletion experiment is recommended.
	Regulation of transcription from Pol II promoter.
	PET9 is linked with ATP/ADP exchange. Therefore, we suspect that YKR046C (PET10) is active in similar function.

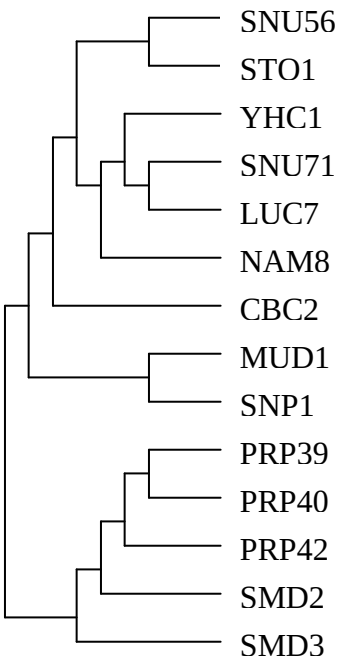
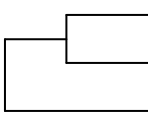
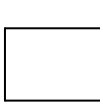
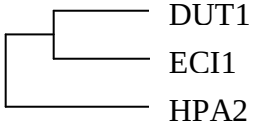
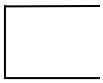
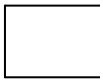
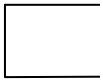
	They form mitochondrial large ribosomal subunit.
	GBP2 is involved in telomeric DNA binding, whereas GBP2 is suspected to play a role in transcription elongation by Pol II. Double mutation would clarify their closer link. They are also weakly linked with HPR1 (DNA-dependent transcription) and MFT1 (protein-mitochondrial targeting) proteins.
	MSS116 is a RNA helicase linked with RNA splicing. YLR432W (IMD3) is IMP dehydrogenase. However, their strong links with other proteins involved with RNA metobolism (such as NOP12) make us to suspect that they are both in that pathway. Double mutation of them may establish the point.
	TFIID complex and related regulators.
	They are both class V myosins involved in endocytosis.
	TIF34, TIF35, NIP1, PRT1, RPG1, TIF5 are involved in translation initiation as part of eIF3 complex. Unannotated protein HCR1 is strongly suspected to be linked with the same process.

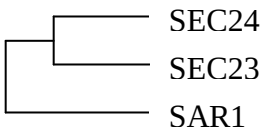
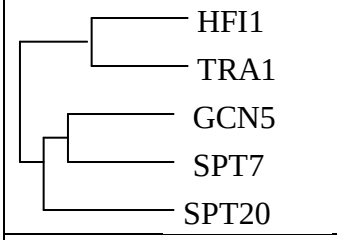
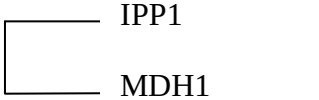
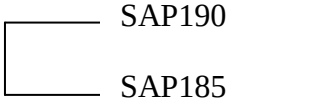
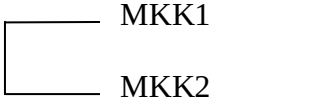
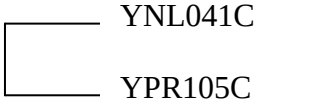
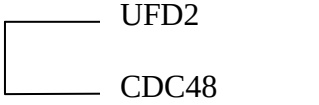
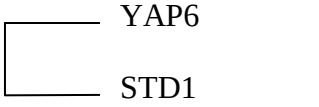
	GTT1 is linked with glutathione metabolism. Therefore, it is suspected that unannotated YEL017W (GTT3) is linked with similar metabolic purpose.
	Parts of Pol III complex.
	CAR2 is involved in amino-acid metabolism. OYE2 is also a NADPH dehydrogenase. Therefore, it is likely that other two yet unannoted proteins are also linked with similar metabolic purposes. Since they are all individually viable, multiple deletion experiments may reveal their functional link.
	RCS8 and ISW1 are involved in chromatin modeling. ITC1 is protein with unknown function.
	Parts of ubiquitin ligase complex involved in G1/S transition of mitotic cell cycle.
	Golgi retention or retrograde transport. Forms associations with proteins in VPS17 module.
	CHD1 is a Pol II transcription elongation factor. FKH1 and ABF1 are transcrition factors related to chromatin silencing at HML and HMR.
	Parts of epsilon DNA polymerase complex, involved in DNA mismatch repair and strand elongation.

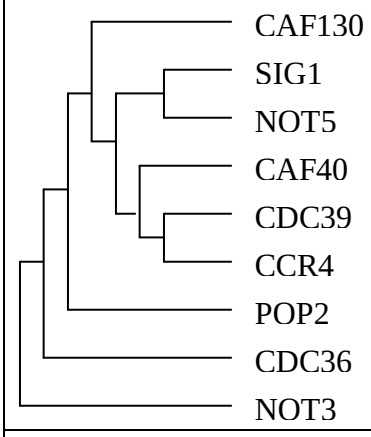
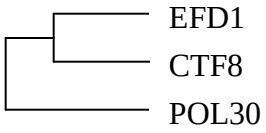
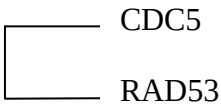
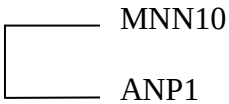
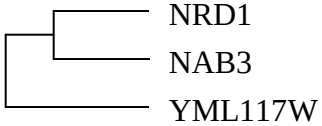
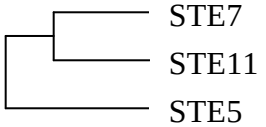
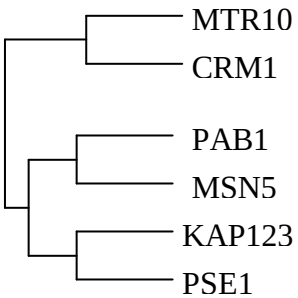
	Mitochondrial small ribosomal subunit.
	They are involved in snoRNA binding, 35S primary transcript processing, processing of 20S pre-rRNA, rRNA modification.
	Signaling proteins involved in establishment of cell polarity, bud growth and shmooing.
	Involved in mRNA splicing in small nuclear ribo-nucleoprotein complex.
	KTR3 is a mannosyltransferase involved in cell-wall synthesis and biogenesis. The whole complex either has similar function or protein-vacuolar targeting.

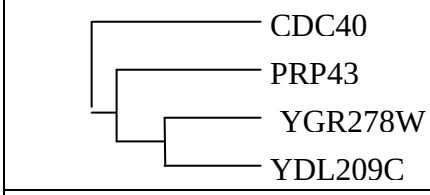
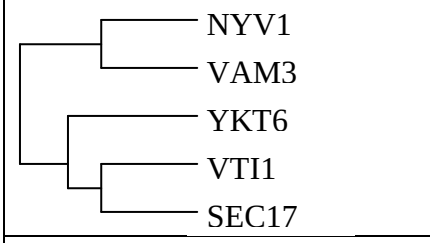
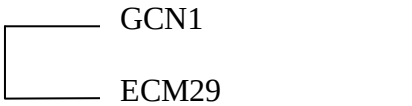
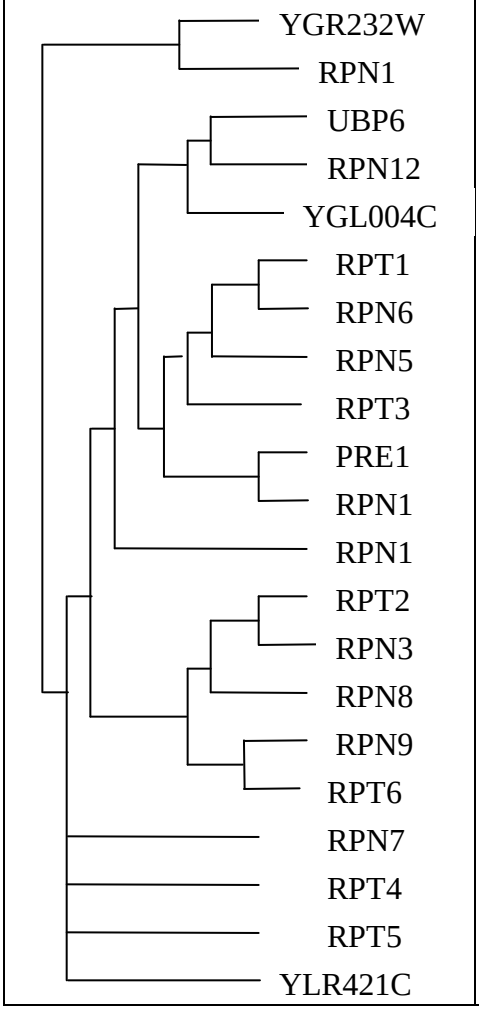
	CDC73/PAF1 complex involved as transcription elongation factor from Pol II promoter.
	Both of them are protein transporters involved in mitochondrial translocation.
	Both proteins have been located with mitochondrion. MHR1 is a transcription regulator involved in mitochondrial genome maintenance, whereas YDR116C is part of mitochondrial large ribosomal subunit.
	Both are involved in mRNA-splicing.
	Both unknown. Based on our study, we suspect these and following other proteins are involved in processing of 27S pre-rRNA ribosomal subunit. They are NOG1, YGR103W, HAS1, CDC95, RLP7, YKR081C, YHR052W, YMR049C, YTM1, NOP2, YDR101C, YOR206W, YNL110C.
	RTS1 is part of protein phosphatase 2A complex. Therefore, it is possible that YGR161C (RTS3) is also involved in similar function.
	ACF2 is involved in actin cytoskeleton organization and biogenesis. Yet unannotated YJR083C (ACF4) is strongly suspected to be linked to the same process.

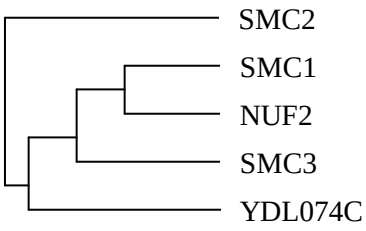
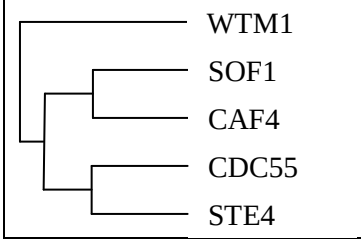
	Mediator complex for transcription from Pol II promoter.
	Anaphase-promoting complex involved in cyclin catabolism, mitotic chromosome segregation and metaphase/anaphase transition.
	Structural constituent of cytoskeleton present in spindle pole body and involved in microtubule nucleation.
	Nucleosome remodeling complex involved in chromatin modeling. Should be probed for double and triple deletion for better understanding.
	Glycogen metabolism.

	Commitment complex and snRNP U1.
	All unannotated proteins possibly involved in biogenesis and transport of ribosome.
	Parts of alpha DNA polymerase-primase complex involved in DNA replication initiation.
	Involved in metabolism.
	Both involved in acetyl-CoA biosynthesis.
	DIM1 is involved in 35S primary transcript processing and rRNA modification. Therefore, it is suspected that YOR145C (DIM2) is involved in similar process.
	SOD1 is related to copper homeostasis, whereas UBA1 is linked with ubiquitin cycle.

	COPII complex.
	SAGA complex linked with chromatin modeling and histone acetylation. Only TRA1 is linked with TRAPP complex. They all take role in transcription control through chromatin modeling.
	MDH1 is a malic enzyme linked with tricarboxylic acid cycle. Therefore, unannotated IPP1, an inorganic diphosphatase is expected to have similar function.
	Involved in G1/S transition of cell cycle. Double mutant grows slowly. Triple mutant with SAP155 is inviable.
	Both are map kinase kinase. Single mutant is viable but double mutants show some defects.
	Parts of Golgi-transport complex and involved in intra-golgi transport.
	Both involved in ubiquitin-dependent protein catabolism.
	YAP6 is a transcription factor, whereas STD1 is involved in signal transduction and regulation of transcription from Pol II promoter. Double mutant should be studied.

	CCR4-NOT complex regulating transcription from Pol II promoter and active in poly-A tail shortening. POP2 is required for glucose derepression.
	CTF8 and POL30 are involved in DNA replication and repair. It is likely that EFD1 (YOR144C) is active in same process.
	They are both protein threonine/tyrosine kinases involved in DNA repair and replcation.
	Parts of mannosyltransferase complex.
	NRD1 is involved in nuclear RNA binding, whereas NAB3 is involved in poly-A binding. Therefore, YML117W (NAB6), an unannotated protein is possibly involved in the same function.
	MAP kinase proteins involved in the signal transduction of mating signal.
	All the proteins except PAB1 are involved in transport of proteins and mRNAs between nucleus and cytoplasm. Only PAB1 is linked with regulation of translation initiation. The reason why PAB1 got associated here with transport proteins is not clear to us.

	Both CDC40 and PRP43 are involved in the splicesome complex and functions as pre-mRNA splicing factors. Therefore, YDL209C and YGR278W are suspected to have similar functions.
	They are all part of SNARE complex and involved in transport between Golgi and other vesicles as well as non-selective vesicle fusions.
	In our analysis, ECM29 shares many pairs with the proteasome complex member proteins. It is linked with cell-wall organization and biogenesis, whereas GCN1 is linked with regulation of translational elongation. Their link is not clear to us.
	All these proteins are parts of proteasome complex linked with proteolysis and peptolysis.

	SMC2, SMC1, NUF2 and SMC3 are involved in chromosome condensation and segregation processes. It strongly suggests that YDL074C (BRE1) is involved in the similar function.
	STE4 is a G-protein GTPase active in signaling during mating. CDC55 is a protein phosphatase. CAF4, WTM1 are other regulatory proteins. SOF1 is part of small nucleolar ribo-nucleoprotein complex. Their link is not clear to us.